CCR-Sequencing Facility Illumina Sequencing Report

Project Information

Principal Investigator:

PI Laboratory Contact:

Bioinformatics Contact:

Project Title

CSAS Order ID:

Samples Total in project: 10

Samples in This Report: 10

Completion of CSAS: yes

Report Date: January

Sequencing Details

Flowcell ID:		Sequence Control:	PhiX
Instrument:	HiSeq2500	Control Result:	Pass
Sequencing Type:	mRNA-seq	Library Protocol:	Illumina TruSeq RNA Protocol RS-930-2001
Read Length:	101 (2x101cycles)	Sequencing Chemistry:	Illumina TruSeq V4.0
		Reference Genome:	Hg19
Multiplexed:	Yes, 5 per lane	Target Region File:	Ensemble v70

Run Comments

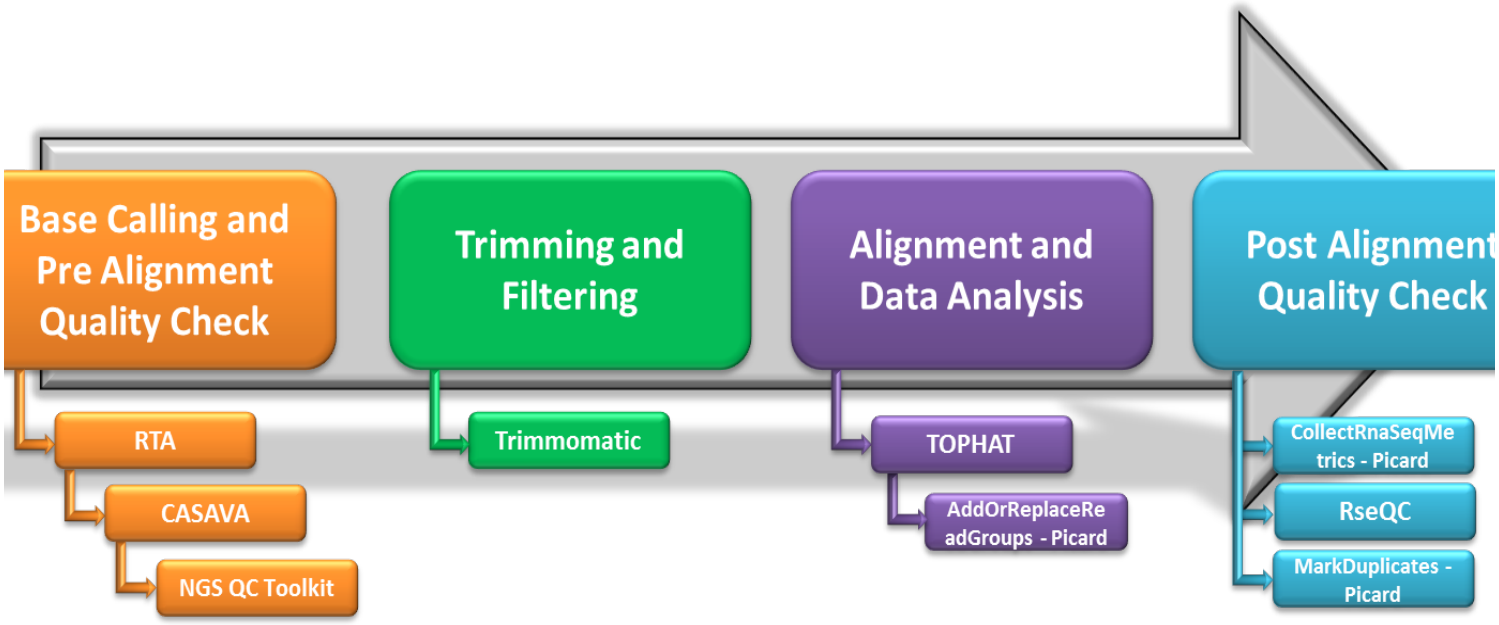
Ten total RNA samples were sequenced five per lane on two lanes of HiSeq2500 with illumina TruSeq V4 chemistry. All sample yields ranges from 74 – 93 million pass filtered reads. Samples have quality with %>= Q30 of above 93%. Sample reads were trimmed adapters and low quality bases using Trimmomatic software and aligned with reference human hg19 genome and ensemble v70 transcripts using Tophat software. All samples align well with the reference human genome above 72% and unique alignment is above 69%.RNA mapping statistics are calculated using Picard software. Sample percent coding bases above 59%, mRNA bases above 60% and ribosomal bases are between 9-17% ribosomal bases. Library complexity is measured by unique fragments in the mapped reads using Picard's Markduplicate utility. Percent non-duplicated reads for all the samples are above 59%.

Note: *Residual samples will be retained up to* ***90 days*** *of the delivery of this report. To avoid shipping charges, please contact SFILLUMINALAB@mail.nih.gov to arrange pickup samples prior to this time.*

Note: *Sequencing data will be available to download for* ***two weeks*** *following delivery of this report. Please download the data files as soon as possible.*

Analysis Workflow

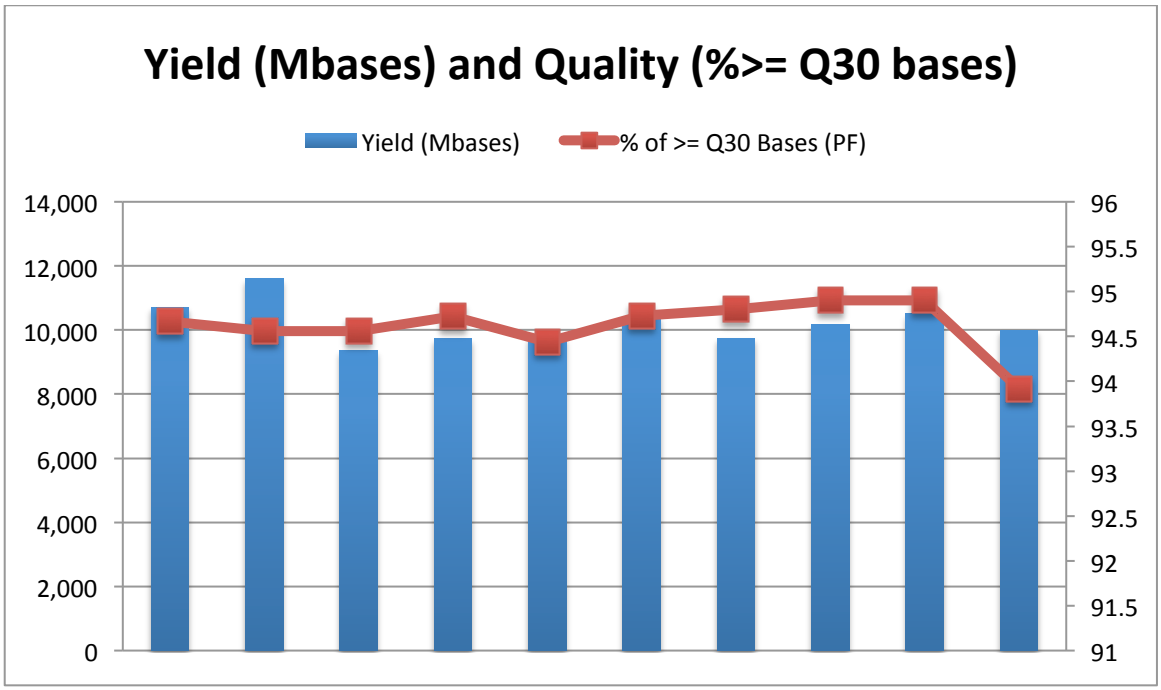
For questions on any aspect of this report please contact ***SFILLUMINABIOINF@mail.nih.gov.***

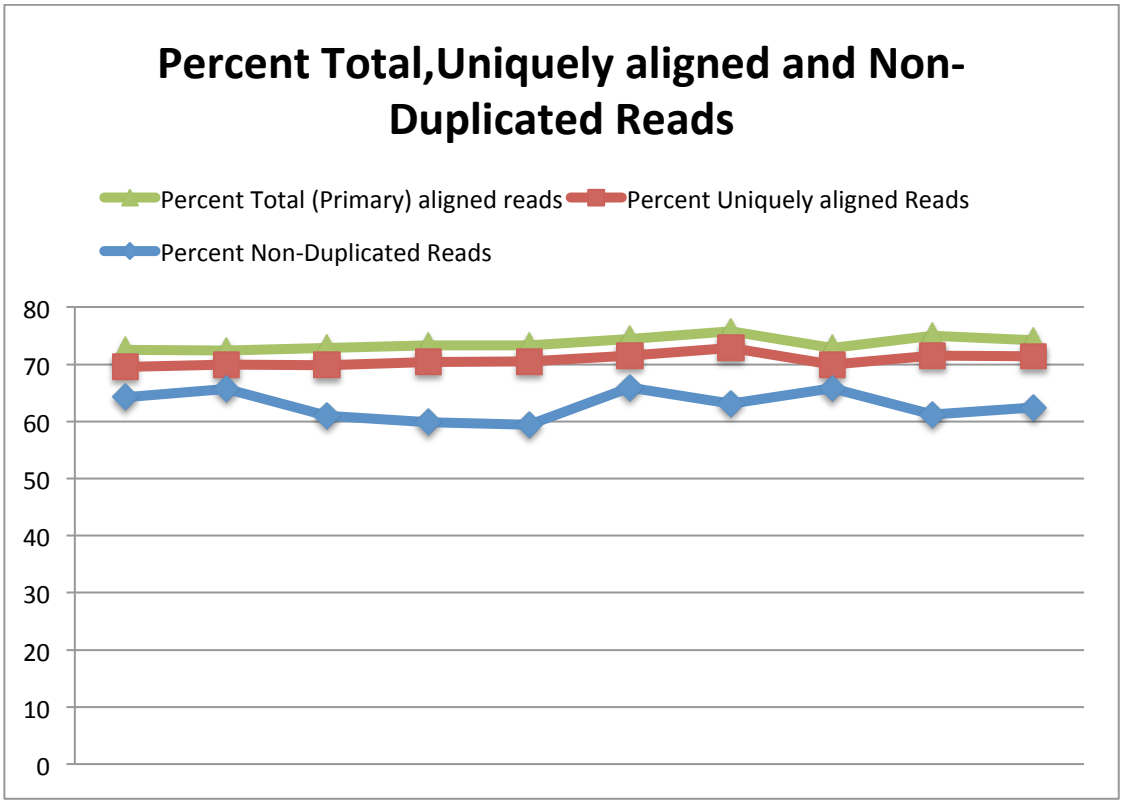


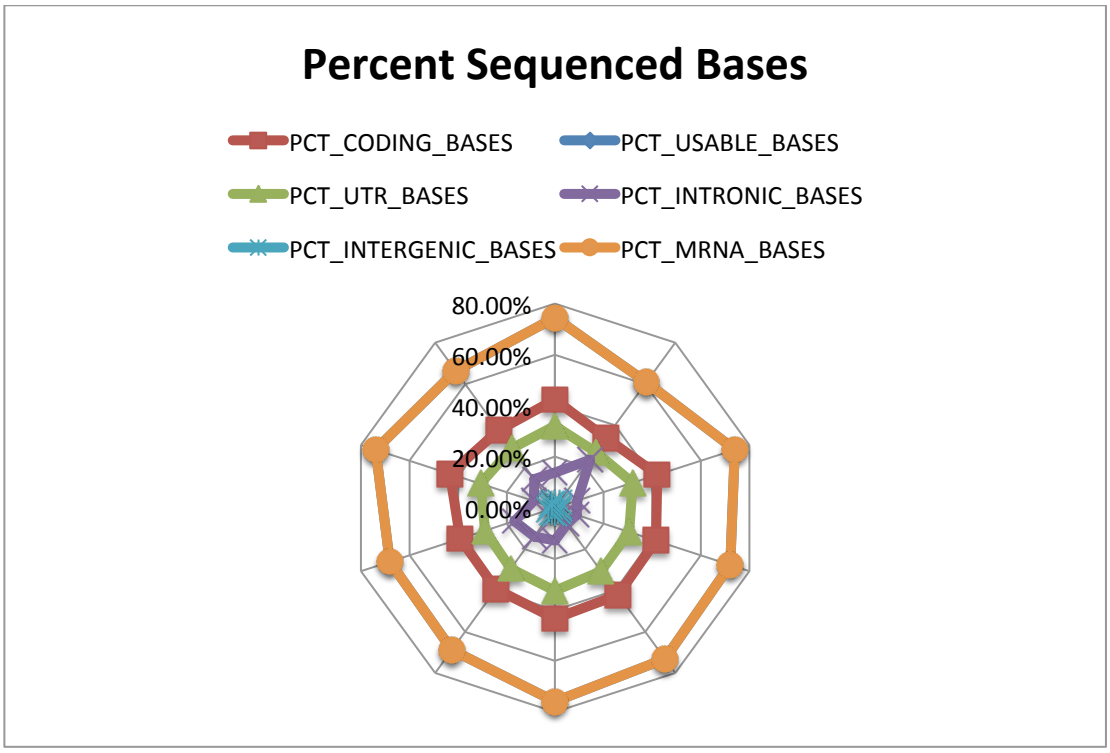
Software and Parameters

Analysis Step	Software	Software Parameters / Notes
Basecalling	RTA 1.18.61	Illumina instrument run time analysis software
Demultiplexing	Casava 1.8.4	Barcode demultiplexed allowing 1 mismatch
Filtering (Adaptor and quality)	Trimmomatic 0.30	PE -threads 16 -phred33 ILLUMINACLIP:TruSeq_and_nextera_adapters.fa:3:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:4:20 MAXINFO:50:0.8 MINLEN:25
Alignment	TopHat v2.0.8	tophat -G hg19_Ensemble_v70.gtf -o ./ -r 10 --mate-std-dev 200 -p 16 read1.fastq read2.fastq
RNAStatistics	Picard 1.84	CollectRnaSeqMetrics.jar REF_FLAT=hg19_Ensembl_v70_refFlat.txt INPUT=sample.bam OUTPUT= RnaSeqMetrics.txt RIBOSOMAL_INTERVALS= ribosome_interval_list.txt STRAND_SPECIFICITY=NONE VALIDATION_STRINGENCY=LENIENT
Duplication Statistics		MarkDuplicates.jar INPUT=sample.bam OUTPUT=sample.MKDUP.bam METRICS_FILE=sample.bam.metric ASSUME_SORTED=true MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000 VALIDATION_STRINGENCY=LENIENT
Insert Size Statistics	RseQC 2.3.5	inner_distance.py –i sample.bam –o ./ -r hg19_Ensembl.bed

Data Statistics


Yield (Mbases) and Quality (%>= Q30 bases)
Yield (Mbases)
% of >= Q30 Bases (PF)
14,000
12,000
10,000
8,000
6,000
4,000
2,000
0
96
95.5
95
94.5
94
93.5
93
92.5
92
91.5
91


Percent Total,Uniquely aligned and Non-Duplicated Reads
Percent Total (Primary) aligned reads
Percent Uniquely aligned Reads
Percent Non-Duplicated Reads
80
70
60
50
40
30
20
10
0



Notes

- **Sample Yield –** The sum of all bases in reads that passed filtering per sample. Indicates the output in million bases (Mb) per lane.
- **% >=Q30 –** The percentage of bases called with an inferred accuracy of 99.9% or above, a measure of basecalling quality.
- **% Total (Primary) Alignment –** The percentage of filtered reads that align to the reference; for mRNA-seq, to the reference genome and the splice junctions. Reads aligning to multiple locations are included in the calculation
- **% Unique Alignment –** The percentage of filtered reads that align uniquely to the reference; for mRNA-Seq, the reference genome and known splice junctions. Reads aligning to multiple locations and abundant sequences are not included in the score.
- **% Non-duplicated Reads –** The percentage of aligned reads with non-redundant start coordinate.
- **% RNA Statistics –** Collect metrics about the alignment of RNA to various functional classes of loci in the genome: coding, intronic, UTR, intergenic, ribosomal. Also determines strand-specificity for strand-specific libraries.

  PCT_RIBOSOMAL_BASES: RIBOSOMAL_BASES / PF_ALIGNED_BASES

  PCT_CODING_BASES: CODING_BASES / PF_ALIGNED_BASES

  PCT_UTR_BASES: UTR_BASES / PF_ALIGNED_BASES

  PCT_INTRONIC_BASES: INTRONIC_BASES / PF_ALIGNED_BASES

  PCT_INTERGENIC_BASES: INTERGENIC_BASES / PF_ALIGNED_BASES

  PCT_MRNA_BASES: PCT_UTR_BASES + PCT_CODING_BASES